# Brain imaging and forecasting: Insights from judgmental model selection

Weiwei Han[a], Xun Wang[b], Fotios Petropoulos[c,*], Jing Wang[d]

[a]*School of Modern Post, Beijing University of Posts and Telecommunications, China*
[b]*Cardiff Business School, Cardiff University, UK*
[c]*School of Management, University of Bath, UK*
[d]*School of Economics and Management, Beihang University, China*

**Abstract**

In this article, we shed light on the differences between two judgmental forecasting approaches for model selection  forecast selection and pattern identification  with regard to their forecasting performance and underlying cognitive processes. We designed a laboratory experiment using real-life time series as stimuli to record subjects' selections as well as their brain activity by means of electroencephalography (EEG). We found that their cognitive load, measured by the amplitude of parietal P300, can be effectively used as a neurological indicator of identification and forecast accuracy. As a result, judgmental forecasting based on pattern identification outperforms forecast selection. Time series with low trendiness and noisiness have low forecasting accuracy because of the high cognitive load induced.

*Keywords:* forecasting, judgment, EEG, laboratory experiment, decision making, cognitive process

## 1. Introduction

Judgment plays a crucial role in formulating forecasts and is unavoidably integrated within the forecasting process. In this study, we investigate the links of cognitive load (cognitive resources required to perform time series forecasting tasks) with cognitive performance

*Correspondence: Fotios Petropoulos, School of Management, University of Bath, Claverton Down, Bath, BA2 7AY, UK.

*Email addresses:* hanweiw@bupt.edu.cn (Weiwei Han), WangX46@cardiff.ac.uk (Xun Wang), f.petropoulos@bath.ac.uk (Fotios Petropoulos), jim08@buaa.edu.cn (Jing Wang)

(ability to recognize the pattern of the time series) and forecasting performance (ability to forecast a future time series) when judgment is used for forecasting tasks.

Although judgment can inform the forecasting function in many different aspects, one area that has received little attention is that of judgmentally selecting between different forecasts or between different forecasting models. In fact, Petropoulos et al. (2018) were the first to the best of our knowledge to empirically compare the performance of judgmental model selection with the performance of statistical/algorithmic selection. They provided evidence that under specific settings, judgment can perform as well as statistical selection. We extended the study by Petropoulos et al. (2018) and attempted to determine the reasons for the effectiveness of specific settings in eliciting human judgment in forecasting. We did so by means of a laboratory experiment in which we combined a standard computerized task with electroencephalography (EEG) and captured the brain activity of the subjects during the experiment. Our findings enhance our understanding of the cognitive process of forecasting and decision making, an enhancement that in turn will complement existing analytic and empirical studies in managerial judgment.

This paper is organized as follows. In Section 2, we briefly review the existing empirical and experimental research on judgmental forecasting, introduce the cognitive load theory and the EEG technique, and lay out our conceptual framework; Section 3 presents the experiment scheme and data analysis procedure; in Section 4 we report the results from the behavioral and EEG analysis; in Section 5 we propose a cognitive model for the judgmental forecasting process and discuss the implications of the results. Section 6 contains our conclusions.

## 2. Literature review

### *2.1. Judgmental forecasting: Empirical and experimental research*

Judgment can be part of the forecasting process in primarily three ways. First, it may be used directly to produce point forecasts without fitting statistical forecasting models on the data and producing statistical forecasts (for example, see: Lawrence et al., 1985; Carbone and Gorr, 1985; Sanders, 1992; Makridakis et al., 1993; O'Connor et al., 1997; Reimers and

Harvey, 2011). Second, judgment may be used to revise/adjust statistical forecasts produced by forecasting software (for example, see: Fildes et al., 2009; Franses and Legerstee, 2011; Trapero et al., 2013; Petropoulos et al., 2016). Third, judgment can be used to select between statistical models (statistical forecasts). Traditionally, this selection task has been done by the software and completed by either information criteria (Hyndman et al., 2002; Hyndman and Khandakar, 2008), cross-validation techniques (Fildes and Petropoulos, 2015), or by selecting the appropriate methods based on a set of rules (Collopy and Armstrong, 1992; Adya et al., 2001). However, a recent study by Petropoulos et al. (2018) suggests that individual judgmental selections can be as good if not better than statistical selection. On top of this, if judgmental aggregation is considered, then judgmental model selection significantly outperforms statistical selection.

The task of eliciting judgment to select between models or forecasts may be implemented in different ways. Petropoulos et al. (2018) considered two such methods. The first consists of simple selection between different options (sets of forecasts derived from different forecasting models). The second involves identification of the applicable time series patterns (trend and seasonality); consequently, the corresponding model is selected. They designed a laboratory experiment to test the efficacy of the two methods. Their results suggest that the second method (pattern identification) performed better overall. Although we can partly attribute the better performance of the second method to its decomposition nature, past studies have not been conclusive about the added-value of decomposition in forecasting (Goodwin and Wright, 1993). Consequently, we need to further explore and better understand the conditions under which each method performs best. In any case, graphical representation of the methods should be preferred. Research has shown that performance is enhanced, especially for trended series, when data are presented in graphs instead of tables (Harvey and Bolger, 1996). Fortunately, this is consistent with the design and development of modern forecasting support systems.

Judgmental forecasting has been shown to also be affected by various time series features. For example, some research has found noise levels as well as the direction and strength of the trend of a series affect the accuracy of judgmental forecasting (Lawrence et al., 2006).

For instance, Harvey et al. (1997) reported that people perform well in identifying positive linear trends compared with no trend. In another study, Thomson et al. (2013) noted that forecasting performance was higher for intermediate trends than for strong ones but judgmental forecasting was superior in upward trends compared with its performance in downward trends. As for the effect of noise, the results from Sanders (1992) suggested that judgment may bring more benefits for low noise series. The decrease in judgmental forecasting performance with an increase in noise was reported as well by O'Connor et al. (1993). However, Sanders and Ritzman (1992) suggested that although data variability might decrease forecasting accuracy, series with a high degree of noise could be better forecast by practitioners' judgment than by statistical methods. Given that the studies focused on either directly producing judgmental forecasts or judgmentally adjusting statistical forecasts, more research is needed to understand how time series features affect cognitive and forecasting performance in the task of manually selecting between statistical models or selecting between forecasts derived from statistical models (but not judgmentally producing or revising forecasts).

To the best of our knowledge, the relationships between cognitive performance and subsequent forecasting performance have not been studied extensively. The only exceptions are the early studies by Eggleton (1982), who discussed "cognitive representation" and the links between forecasting accuracy and correct assessment of the underlying process of data generation. However, we believe that such relationships should be explored further and explicitly linked with judgmental model selection.

### *2.2. Cognitive theory and EEG*

We have attempted to link the performance of judgmental forecasting with its underlying cognitive process  more specifically, the cognitive load induced while performing forecasting tasks. Cognitive load can be roughly defined as the mental effort exerted in response to cognitive tasks. It plays a crucial role in correctly recognizing the patterns of the time series and generating (or selecting) forecasts. Typical forecasting tasks involve two kinds of cognitive load: intrinsic and extraneous (Choi et al., 2014). The complexity of the task

influences the intrinsic load; the presentation of the task causes the extraneous load. In forecasting tasks, these two kinds of loads can be attributed, respectively, to the differences in time series features and task settings.

Three different methods are currently used to measure cognitive load: task performance and the subjective and objective methods. The task performance method uses the performance of the decision making (e.g., forecasting accuracy or response time) as an indirect indicator. However, because it is an *a posteriori* measure, it cannot be derived until after a decision has been made. Subjective methods that rate the perceived difficulty of a task by surveying research participants have two disadvantages. First, is the difficulty of using a universal subjective rating scale to distinguish between different types of cognitive load. Second, the measurement is taken after the decision activities; thus, the measures fails to track actual cognitive load during the decision process.

In comparison, as Dirican and Göktürk (2011) point out, the psycho-physiological approach can offer a measurement of cognitive load that is objective, sensitive to different cognitive processes, and does not obstruct procedures while maintaining implicitness and continuity. The EEG technique especially adds neuro-physiological visualization to the cognitive process, thus detecting subtle fluctuations instantaneously that other measurements often miss.

The EEG measures brain activity at the scalp level by attaching numerous noninvasive electric detectors onto the scalp and recording the changes in electric potential at these locations (Kenning and Plassmann, 2005). It is one of the most widely applied techniques to reveal perceptual and cognitive activities in the brain. In cognitive tasks, a change in electric potential after presentation of a stimulus is termed the event related potential (ERP). In specific brain regions (see Figure 1 for an overview), positive and negative polarities in the ERP can be found that are referred to as ERP components. The amplitude of these components is often used to indicate the existence or nonexistence of cognitive functions evoked in the brain.

We looked at three ERP components that are closely related to the decision-making

process: P300 [1] in the parietal area, which is associated with the attention and cognitive load (Donchin, 1981); N270 in the frontal area, which is associated with matching tasks (Wang et al., 2000); and the late positive component (LPC) in the frontal area, which reflects short-term memory retrieval (Düzel et al., 1999). The strength of these ERP components indicates the activation of the related brain function. Here "strong" and "weak" refer to the absolute amplitude in the ERP component. For instance, the value of a P300 component in one experiment is said to be stronger than the other if its amplitude (measured by either its maximum or average over a time window) is higher. On the other hand, an N270 is stronger if its value is lower. The timing and associated brain area of these components are shown in Figure 2. We focused on two electrodes in this research, namely Fz and Pz (locations labeled in Figure 1), as representatives of the frontal and parietal areas.

### *2.3. Conceptual framework*

Figure 3 presents the framework and relationships investigated in this study. A key proposition in this research is that judgmental forecasting accuracy (forecasting performance) is affected by the ability to recognize the patterns of the time series (cognitive performance). This recognition in turn is constrained by the cognitive load, which is detectable in the EEG. Hence, we aim to explore the relationships of cognitive performance with cognitive load and forecasting performance as moderated by forecasting task settings and time series features, including trendiness, noisiness, and the direction of the trend.

## 3. Experimental design

### *3.1. Subjects*

Forty subjects (19 males and 21 females) ranging in age from 21 to 28 years (with a mean age of 23.675 years, S.D. = 2.499) participated in our study. All subjects were

[1]Many ERP components are named by their polarity (positive/negative) and latency in time, e.g., P300 stands for a positive component appearing at around 300ms after the presentation of stimulus. The stated latencies for ERP components are often highly variable. For example, the P300 component may exhibit a peak anywhere between 250ms and 500ms, depending on stimulus type, task conditions, subject's age, and other factors.

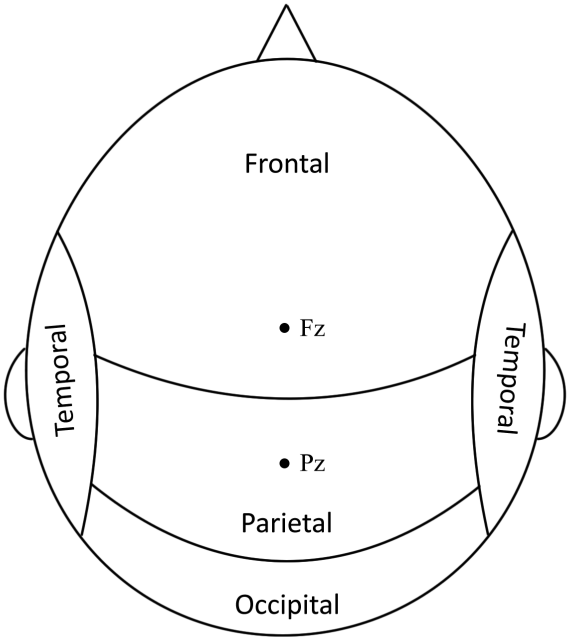


Figure 1: Sketch of brain cortices and electrodes used in analysis. Frontal, temporal, parietal, and occipital are brain cortices. Fz and Pz are the electrodes.

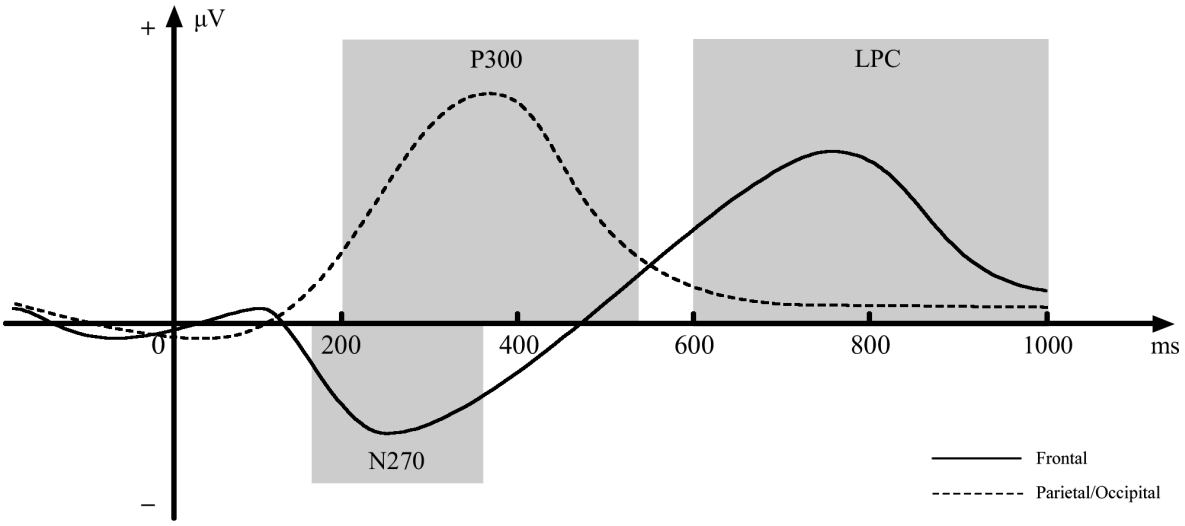


Figure 2: Sketch of ERP components.

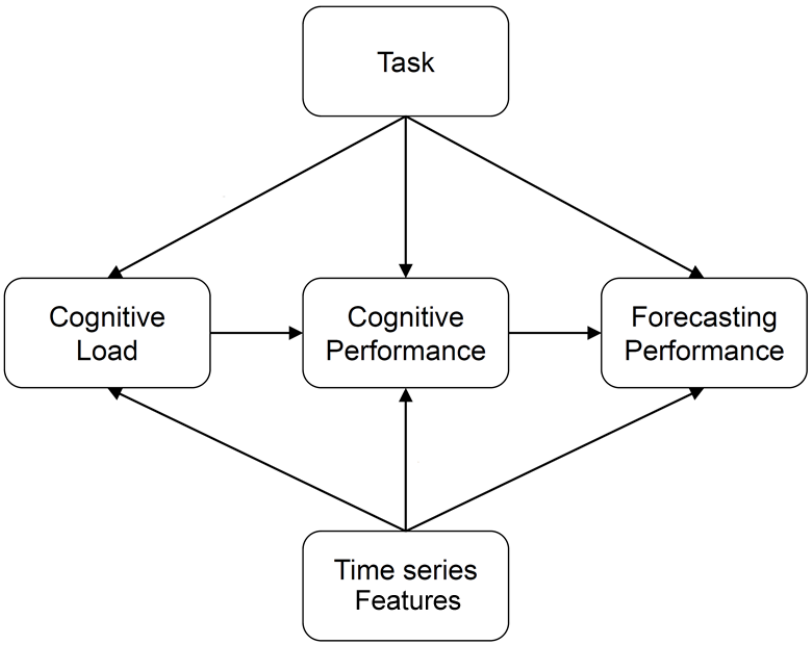


Figure 3: Conceptual framework of this study.

Chinese students studying at the School of Economics and Management, Beihang University. Their native language was Chinese. They were right-handed, and none reported any history of neurological or mental abnormalities. In addition, all were enrolled in a production and operation management course and were familiar with the basic concepts of demand forecasting.

### *3.2. Time series*

The real demand data was drawn from a data set that was a subset of the M3-competition data (Makridakis and Hibon, 2000). More specifically, we have focused on yearly data in which seasonality was inapplicable (as could be the case with quarterly or monthly data). We filtered the 100 longest time series. This provided series with at least 46 observations for both the in-sample and out-of-sample periods. Time series longer than 46 were truncated so that we achieved consistency across the trials. As in the original M3-competition settings, we opted for a split of 40 observations in the in-sample period and 6 in the out-of-sample period (i.e., the forecast horizon equals 6).

We produced forecasts using two models, namely Simple Exponential Smoothing (SES,

a forecasting model suitable for nontrended data) and Holt's Exponential Smoothing (HES, a forecasting model suitable for trended data). Appendix A contains the mathematical formulations for SES and HES. Forecasts were produced using the `ets()` function in the `forecast` package of the R statistical software. The overarching task was selecting between the two models/sets of forecasts based on the in-sample data (the first 40 observations) before the out-of-sample data (last 6 observations) became available.

We used the mean absolute error (MAE) of the out-of-sample period to decide whether a selection was accurate. For the SES and HES forecasts of a time series, the MAE (denoted as $\text{MAE}_{\text{SES}}$ and $\text{MAE}_{\text{HES}}$, respectively) was calculated as $\sum_{i=1}^{h} |y_{n+i} - f_{n+i}|/h$, where $h$ is the forecast horizon ($h = 6$), $n$ is the length of the in-sample data ($n = 40$), $y_t$ and $f_t$ are the actual and the forecast at period $t$, respectively. A selection was accurate in *forecasting* if it corresponded with the forecast with lower out-of-sample MAE, or $A_{\text{MAE}} = I[\text{MAE} = \min(\text{MAE}_{\text{SES}}, \text{MAE}_{\text{HES}})]$ where $A_{\text{MAE}}$ was the accuracy of the forecast and $I(\cdot)$ was the indicator function that translates Boolean values to binary ones (i.e., TRUE becomes 1, and FALSE becomes 0). The $A_{\text{MAE}}$ can be averaged across series and subjects to measure average *forecasting performance*. Note that of the 100 time series used for this research, SES outperformed HES with regard to the out-of-sample MAE in 51 of these series.

We also considered how well subjects correctly detected trends in the in-sample set of the time series. Both SES and HES were fitted for the in-sample of 40 observations, and the model with the minimum Akaike's Information Criterion (AIC) indicated whether the time series trended. The AIC considers the maximum likelihood of each model as penalized by its complexity. As such, the AIC is more suitable for selecting between the in-sample fits of different forecasting models and for testing such selections before the future data becomes available because the MAE or the mean squared error may lead to over-fitting. Moreover, information criteria in general are widely suggested for automatic model selection (see for example: Hyndman et al., 2002; Hyndman and Khandakar, 2008).

We suggest that the judgmental selections, as evaluated by the in-sample AIC, can be used as a proxy to measure *cognitive performance*. Similar to the $A_{\text{MAE}}$, a judgmental selection is accurate in *cognition* if it is in line with the model with the lower AIC, or

$A_{\text{AIC}} = I[\text{AIC} = \min(\text{AIC}_{\text{SES}}, \text{AIC}_{\text{HES}})]$ in which $A_{\text{AIC}}$ is the accuracy of cognition. Similar to forecasting performance, cognitive performance can be averaged across subjects and series. For the time series considered in our study, 58% of them prefer SES over HES based on the in-sample AIC. At the same time, the AIC and MAE preferences coincide in 59% of the series. Table 1 shows how these populations are further divided.

Table 1: Number of time series identified as optimal based on AIC selection versus performing better on the out-of-sample data.

| | | Preferred by MAE | | Total |
|---|---|---|---|---|
| | | SES | HES | |
| Preferred by AIC | SES | 34 | 24 | 58 |
| | HES | 17 | 25 | 42 |
| Total | | 51 | 49 | 100 |

To measure the effects of cognitive and forecasting performance on time series features, we introduced two continuous variables and one categorical variable to measure (i) the degree of trendiness (i.e., strength of trend), (ii) level of noise in data, and (iii) direction of the ongoing trend. Note that the data used in this experiment were not randomly generated by prespecified processes, which necessitated that we define these three variables here.

We devised a trend variable u as an objective and continuous measure of the trendiness of each time series. This variable was calculated as the absolute sum of the rolling HES trend terms of the 0-1 standardized time series. Mathematically, the 0-1 standardized series $\tilde{y}$ was calculated as

$$\tilde{y}_t = \frac{y_t - \min(y)}{\max(y) - \min(y)}.$$

We have $\max(\tilde{y}) = 1$ and $\min(\tilde{y}) = 0$. Such a transform maintains the exact shape of the time series but does not affect the experiment because the time series plots as stimuli are unscaled in the experiment. Next, we fitted the HES model to $\{\tilde{y}_1, \cdots, \tilde{y}_t\}$ and denoted the smoothed trend component as $\tilde{b}_t$. The trend variable $u$ was calculated as

$$u = \frac{1}{n} \sum_{t=1}^{n} |\tilde{b}_t|$$

where $n = 40$ was the length of in-sample.

We contend that $u$ is an appropriate measure of trendiness for real-time series, with a small $u$ indicating low trendiness and a large $u$ indicating high trendiness. This is because (i) $u$ can account for both positive and negative trends because of the absolution; (ii) for stationary time series in which the trend parameter $\beta$ is zero, $\tilde{b}_t$ approaches zero as $t$ increases, thus $u$ tends to be a constant independent of $t$; (iii) $u$ shows a significant distinction with respect to the AIC preference. Specifically, $\mathrm{E}\{u|\mathrm{AIC}_{\mathrm{SES}} < \mathrm{AIC}_{\mathrm{HES}}\} = 0.662$ and $\mathrm{E}\{u|\mathrm{AIC}_{\mathrm{SES}} > \mathrm{AIC}_{\mathrm{HES}}\} = 1.034$, $p < 10^{-6}$.

We measured the noise level by the standard deviation of the standardized data (or the coefficient of variation of the original data) once level changes (trends) were removed. For that purpose, we used the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test to first test each standardized series $\tilde{y}$ for stationarity. If the null hypothesis of stationarity was rejected, then first-order differencing was applied. We repeated this process until the data became stationary, as denoted by $\tilde{y}_s$. Note that for some series no differencing was required. Consequently, the level of noise $v$ was calculated as the sample standard deviation of $\tilde{y}_s$ or

$$v = \mathrm{std}(\tilde{y}_s)$$

Finally, we measured the direction of the ongoing trend based on the signs of the last $\tilde{b}$ value, $\tilde{b}_n$ (note that the signs of values of vectors $b$ and $\tilde{b}$ are identical). We assumed that a series exhibits an upwards trend if $\tilde{b}_n > 0$; otherwise, we assumed that a series exhibits a downward trend. In total, 76% of the series trended upward.

### *3.3. Experiment Paradigm*

The experiment was conducted in the Behavioral and Human Factors Laboratory in School of Economics and Management, Beihang University. During the experiment, the subjects were seated in a comfortable armchair in a sound-attenuated room. They were instructed to avoid blinking or moving their eyes and to place their fingers on the keyboard, the left index finger on the F key, and the right index finger on the J key. Stimuli were

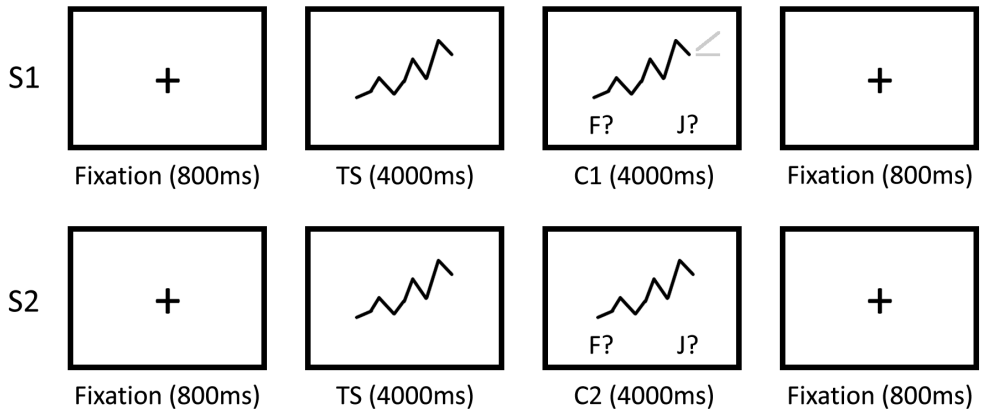


Figure 4: An illustration of the trial structure.

presented on a 20-inch computer monitor located one meter in front of the subjects (with a 10-degree visual angle). The monitor was controlled by a personal computer running E-Prime 2.0 (Psychology Software Tools, Inc.) and Net Station 5.4 (Electrical Geodesics, Inc.).

The experiment consisted of two rounds, each corresponding to one of the two tasks (S1 and S2) and contained 50 trials. To eliminate a temporal effect, subjects were randomly divided into two groups. The first group performed S1 and S2 in sequence. The order was reversed for the second group. At the beginning of each task, subjects viewed a briefing slide that introduced the procedure and expectations in the subjects' native language (Chinese). Subjects were permitted a 10-minute break between the two tasks. In each task, the subject needed to conduct 50 randomly selected trials from the pool of 100 time series. The exhibition sequence was also randomized. The frequency with which each series was presented also closely followed a uniform distribution. Each trial had an exhibition phase (TS) and a selection phase (C1/C2). The exact process of each trial within a task and its phases are detailed below. Figure 4 also illustrates the process.

The S1 task asked a subject to choose a forecast from two sets of forecasts, respectively, from nontrended (SES) and trended (HES) models, given the in-sample plot. The choice of one forecast line or the other directly reflects on the choice of the respective method. Each trial of S1 began with a fixation phase (a red plus sign "+" appearing in the center of the

screen) for 800ms. The exhibition phase of the time series (TS) came next. This phase of S1 graphically displayed the 40 in-sample data points. The tickers and legends on the axis of the plots were removed to minimize interference. The TS phase lasted for 4000ms. The C1 phase added to the existing graph the point forecasts for the next six periods generated by SES (blue forecast line) and HES (green forecast line). Subjects were required in the C1 phase to decide which forecast (blue line or green line) was better and then to press the corresponding key on the keyboard (F for blue and J for green). The next trial began after the subject's response or after 4000ms from the beginning of C1 if the subject did not respond.

The S2 task involved pattern recognition in which the subject was asked to answer whether a trend occurred in the plot of a given time series. The identification of a trend (or lack thereof) can be subsequently translated to selection of one of the forecasting models (SES if no trend is identified, HES if one is identified) and, indirectly, to its forecasts. Similar to S1, each trial began with the fixation and TS phases, which were shown for 800ms and 4000ms respectively. Unlike in S1, the TS phase was followed by the C2 phase in which a subject used the F and J keys to indicate his or her decision on whether the series exhibited a trend. Similar to S1, the next trial began after a subject's response or after 4000ms from the beginning of C2 when there was no response.

Subjects did not receive feedback after each trial or task so we do not need to consider the learning effect. Trials with no response were marked as "inaccurate" in both tasks but their EEG results were still used in the analysis. The only EEG results excluded were those with high noise levels and those in which the software malfunctioned. After those exclusions, the analysis used 3,722 of 4,000 trials. Of those, 1,832 trials were in SI and 1,890 in S2.

## 4. Results

We used EEGLAB 14.1.1b in a MATLAB R2015a environment and R statistical software for most of the behavioral and ERP analysis for cognitive load, cognitive performance, and forecasting performance. A discussion of our results first requires an introduction of the categorization approach we adopted. Analysis of EEG results often requires multiple trials

to average out noises unrelated to the cognitive process such as blinking. For instance, roughly 1,800 trials were averaged to compare S1 and S2. However, the real effect of time series features that are measured continuously often cannot be seen because too few trials were used in averaging. Consequently, in discussing the features of time series and EEG signals, we have divided all trials into three categories based on the following approach: We first sorted all trials based on the values of $u$ and $v$ of the time series, then we divided these trials into three subsets of approximately the same size. Repeated trials with the same values of $u$ and $v$ are categorized together. These three categories represent low, medium, and high trendiness or noisiness. For the direction of the trend defined in section 3, the number of categories is naturally two (positive or negative).

We examined the impact of both experimental tasks (S1/S2) and the time series characteristics (trendiness, noisiness, and trend direction) on cognitive load, which is measured by the P300 component (Table 2). Specifically, we computed the average electric potential at location Pz between 200ms and 600ms.

Table 2: P300 level, AIC accuracy, and response time

| | | S1 | | | S2 | | |
|---|---|---|---|---|---|---|---|
| | | $A_{\text{AIC}}$ | $A_{\text{MAE}}$ | P300 | $A_{\text{AIC}}$ | $A_{\text{MAE}}$ | P300 |
| Trendiness | Low ($\bar{u} = 0.41$) | 0.410 | 0.490 | 4.347 | 0.501 | 0.469 | 3.348 |
| | Medium ($\bar{u} = 0.83$) | 0.575 | 0.462 | 2.939 | 0.661 | 0.526 | 2.240 |
| | High ($\bar{u} = 1.21$) | 0.655 | 0.556 | 2.917 | 0.718 | 0.603 | −0.247 |
| Noisiness | Low ($\bar{v} = 0.04$) | 0.735 | 0.574 | 1.953 | 0.870 | 0.653 | 0.232 |
| | Medium ($\bar{v} = 0.10$) | 0.477 | 0.477 | 4.253 | 0.402 | 0.455 | 3.348 |
| | High ($\bar{v} = 0.23$) | 0.439 | 0.462 | 3.924 | 0.609 | 0.493 | 1.612 |
| Direction | Negative | 0.415 | 0.435 | 5.432 | 0.595 | 0.480 | 1.959 |
| | Positive | 0.595 | 0.534 | 4.030 | 0.651 | 0.552 | 2.488 |
| Average | | 0.552 | 0.510 | 4.366 | 0.637 | 0.535 | 2.359 |

Comparing the task settings shows that subjects performed worse in S1 than in S2 in terms of both pattern recognition (0.552 vs. 0.637 in AIC accuracy) and forecast selection (0.510 vs. 0.535 in MAE accuracy). This is accompanied by a higher P300 in S1 than in S2 (4.366 vs. 2.359). Assessing trendiness and noisiness, we found that subjects perform better when the time series is more trended or less noisy (except between medium and high

noisiness in S2). This is true for both AIC and MAE accuracy. The direction of the trend also has a significant impact on AIC accuracy. In S1, AIC accuracy is 0.435 for the negative trend and 0.534 for the positive; in S2, they are 0.595 and 0.651, respectively. In almost all cases, a high P300 level corresponds to low AIC accuracy (except between negative and positive directions in S2). The differences between categories are all significant at a 0.05 level.

From the above observations, we can derive the following general findings. The AIC and MAE accuracy tends to increase or decrease simultaneously, and high P300 is associated with low AIC and MAE accuracy. However, the contrast in MAE accuracy between categories is weaker than with AIC accuracy. As for moderating variables, S2 is beneficial for achieving higher accuracy, and time series with higher trendiness, lower noisiness, and positive trend directions are easier to forecast.

We further demonstrated the robustness of the results by a correlation analysis between the time series features ($u$ and $v$) and AIC and MAE accuracy by increasing to 10 the number of categories (Table 3). The cognitive load variable is omitted because the number of trials in each category is insufficient for an EEG analysis. We observed a strong correlation between $A_{\text{AIC}}$ and $A_{\text{MAE}}$, which is dictated by the characteristics of the data set. In all cases, a positive (negative) correlation occurs between trendiness (noisiness) and AIC/MAE accuracy. Moreover, such correlations are stronger in S1 than in S2. Lastly, the correlation between the features and $A_{\text{MAE}}$ is generally weaker than that between the features and $A_{\text{AIC}}$.

Table 3: Correlation between time series features, AIC accuracy, and MAE accuracy

| Categorized by | Task | $(A_{\text{AIC}}, A_{\text{MAE}})$ | $(u, A_{\text{AIC}})$ | $(u, A_{\text{MAE}})$ | $(v, A_{\text{AIC}})$ | $(v, A_{\text{MAE}})$ |
|---|---|---|---|---|---|---|
| Trendiness | S1 | 0.4853 | 0.6272 | 0.3013 | - | - |
| | S2 | 0.6314 | 0.3324 | 0.2518 | - | - |
| Noisiness | S1 | 0.7654 | - | - | $-0.6397$ | $-0.5205$ |
| | S2 | 0.7578 | - | - | $-0.2905$ | $-0.4265$ |

## 5. Discussion

### *5.1. Cognitive processes in S1 and S2*

Section 4 shows that, exposed to the same time series plots, subjects reacted differently to two tasks that appeared as distinctions in cognitive and forecasting accuracy. We found that cognitive load, represented as the amplitude P300 component (Figure 5a), is a crucial factor of forecasting accuracy. However, it remains unclear what caused the difference in cognitive load between tasks. Through further ERP analysis, we propose in this subsection that the distinctions in cognitive load between the two tasks can be explained by the differences in the cognitive processes involved.

The asymmetry between the left and right hemispheres in the TS phase can be seen in S1, but not in S2, as shown by the topographical map in Figure 5b. (A topographical map color-labels the level of the electric potential at the locations of all electrodes at a given point in time.) This is clear evidence that working memory storage occurs in S1 but is absent in S2 (Tulving et al., 1994). On the other hand, in the C1/C2 phases, LPC in the frontal area at around 600-1000ms (Figure 5c), which is related to the retrieval of working memory, is more prominent in S1 than in S2. These observations reveal a distinctive difference between S1 and S2 in memory processing. In S1, memories involved with decision making are predominantly short-term; this means that for each individual trial, subjects' memories of the in-sample time series and their personal judgment have to be stored and retrieved. Moreover, this retrieval can only be performed during C1 to permit the subjects to match their judgment with the options. Instead, we did not detect a high relevance between working memory and decision making in S2. In fact, the only memory-related activities involved in S2 are storage, retrieval of categorical decisions (F or J) in working memory, and retrieval of trend criteria from long-term memory that is not dependent on an individual time series. The fact that the time series features do not affect either topographic asymmetry or LPC amplitude serves to further verify these association. This lack of effect also indicates that these associations are induced by the specific task settings and corresponding cognitive processes.

Frontal N270 is induced by the matching function (Wang et al., 2000). Here the match/mismatch

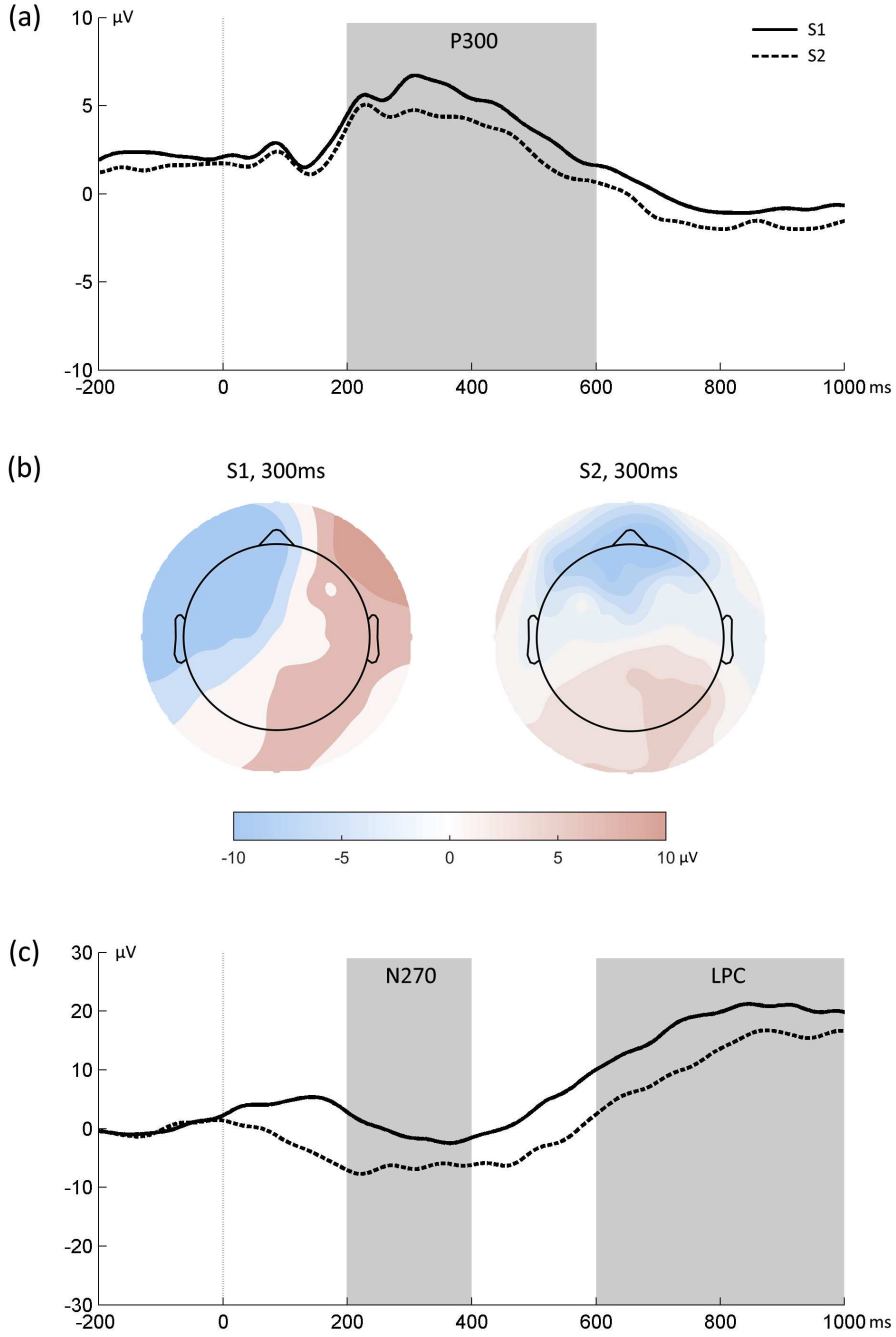


Figure 5: EEG and topographic plots of the brain response in S1 and S2. (a) P300 in TS phase. (b) Topographic map in TS phase at 300ms. (c) N270 and LPC in C1/C2 phases.

can either be between different stimuli or between memory and stimulus. The amplitude of N270 was found to increase with the task-relevant mismatch but decrease with the task irrelevant mismatch (Wang et al., 2004). In this study, a weaker N270 occurs in S1-C1, where subjects try to match their judgment with one of the two forecast lines but cannot always find a perfect match. However, this is attenuated by two task-irrelevant mismatches (the appearance of forecast lines and options). In contrast, in S2, the subjects can always find a match between their judgment and the two options. Also, there is only one task irrelevant mismatch (the appearance of options). The differences in N270 and LPC between the two tasks are statistically significant at a 0.001 level.

Figure 6 summarizes the cognitive processes that occur in S1 and S2. This is in line with the general cognitive processes described in the widely adopted CAT-R cognitive model (Adaptive Control of Thought - Rational, Anderson et al., 2004). In S1, the subjects first form their own judgment of the forecast in the TS phase and then encode the judgment in their working memory. After the options were shown in S1-C1, the subjects retrieved the pattern of the judgment from their working memory and tried to match the judgment with one of the two options, although there was no guarantee that they would find an option that perfectly matched the judgment. In S2, the process involved little encoding of working memory. Instead, the process evoked long-term memory retrieval in which the subjects recalled the characteristics of trended/nontrended series, a step independent of the individual trials. The subjects were able to make the decision in the TS phase and simply selected the corresponding option in the C2 phase.

In general, S1 required more cognitive resources, in terms of frequent storage and retrieval of working memory, as well as in matching judgments and options. Further studies have shown that a choice made in S1 is not statistically associated with the trend of the time series, indicating that subjects may use heuristics to make decisions. Thus, we concluded that higher cognitive load leads to irrational decisions, hence inferior performance in pattern identification and low forecasting accuracy.

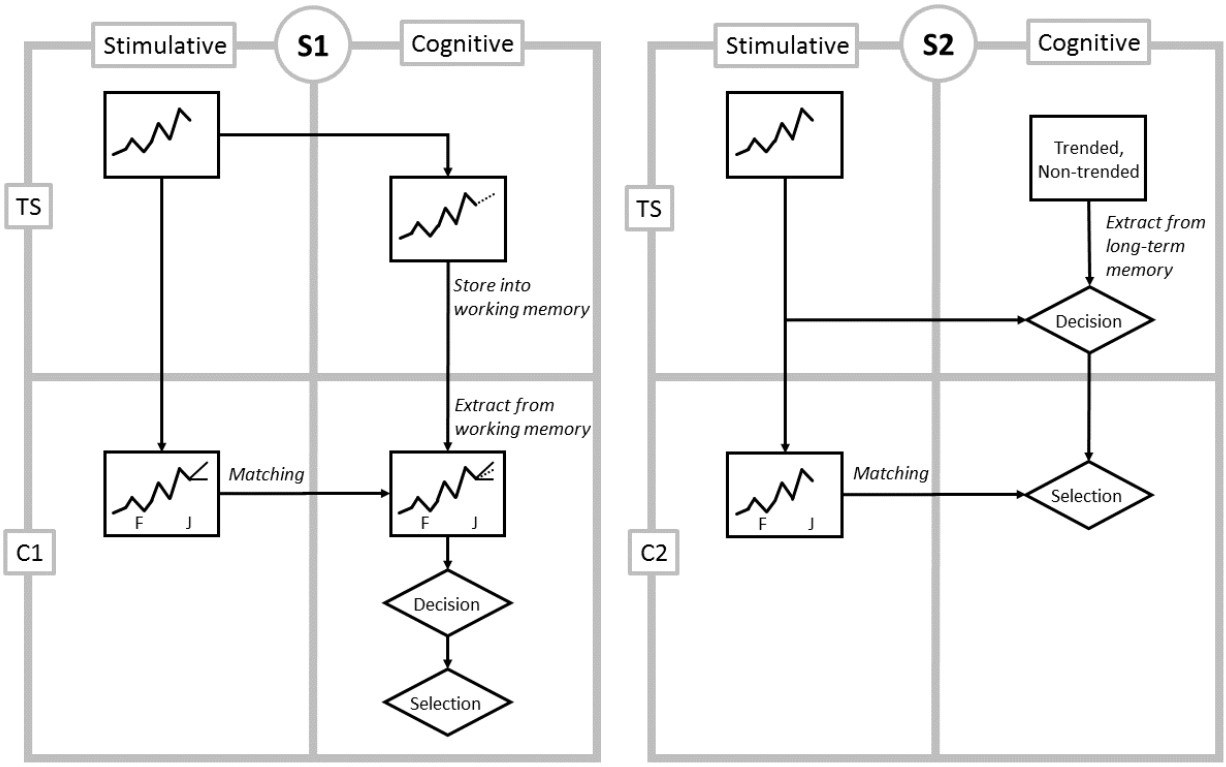


Figure 6: A visual summary of the differences in the cognitive processes for the two tasks.

*5.2. Implications of the findings*

The results from the analysis of our data yield some useful insights that can be translated into recommendations for the design of forecasting support systems.

First and foremost, we found a significant negative relationship between cognitive load and cognitive performance as well as a significant positive relationship between cognitive performance and forecasting performance. The cognitive load of the forecasting task is affected not only by the features of the time series but also by the task settings. These two factors may correspond to intrinsic and extraneous load as the theory suggests. Thus, we conclude that the performance of judgment in forecasting tasks can be affected not only by the complexity of the task but potentially also by the complexity of the systems used to support decision making. For that reason, we should aim for simplified processes and tasks and support managerial judgment to the greatest degree possible so that cognitive performance is maximized.

We confirmed the findings by Petropoulos et al. (2018) that the task (S1 or S2) affects forecasting performance and especially that subjects perform better in S2 than in S1. Even if the differences in forecasting performance are not statistically significant, the differences between S1 and S2 are statistically significant when cognitive load and cognitive performance are considered. This suggests that even if the benefits in the out-of-sample accuracy of S2 over S1 are statistically marginal, S2 should nevertheless be preferred over S1. This preference is because S2 saves valuable resources by reaching its performance levels with the expenditure of significantly less cognitive load and managerial time. Consequently, we suggest that the design of future forecasting support systems should follow a model-build strategy (users can decide whether to include specific time series components such as trend and seasonality with such decisions subsequently translated into the respective models) instead of simply selecting between alternative forecasting models. We believe that our results provide evidence toward the positive effects of the decomposition strategies.

Our study also complements the existing literature through its exploration of the effects of time series features when judgment is used in the forecasting process. We found that trends has a positive relationship with cognitive load and cognitive and forecasting perfor-

mance. Moreover, such a relationship becomes less significant as we move from cognitive load to cognitive performance to forecasting performance. We have reinforced the findings by Thomson et al. (2013) that trend direction significantly affects forecasting performance, with positive trends resulting in higher performance, We also have shown that such a relationship exists for both cognitive load and cognitive performance. Our results also show that judgment works best for series with low noise, confirming the findings by Sanders (1992) and O'Connor et al. (1993).

## 6. Concluding remarks

Judgmental forecasting has long been an intriguing topic for academics and practitioners in business management, especially when statistical forecasting methods are inapplicable or unsuitable. The cognitive process of human forecasters plays a central role in judgmental forecasting, however, very few researchers have devoted their attention to how humans actually reach their judgments in forecasting tasks. This experimental research for the first time provides evidence of links between forecasting accuracy and neuro-physiological activities.

We showed that high cognitive load leads to low cognition accuracy, thereby low forecasting accuracy. This explains the effect caused by task settings and time series features. Significant differences in brain activity have been observed between forecast selection and pattern identification. Distinctive cognitive functions that these two tasks trigger may account for these differences. We have also found that cognitive accuracy and forecasting accuracy increase with trendiness, decrease with noisiness, and are higher when the trend is in a positive direction. All relationships were found to fit with the predictions of cognitive load.

This finding may lead to the following important insight. When performing judgmental forecasting tasks, human forecasters should position their intervention sooner rather than later, a positioning that could have significant implications for the design and development of forecasting support systems. Although statistical and algorithmic approaches have dominated selection of automatic forecasting models, we foresee that a carefully designed

brain-computer interface (Valeriani et al., 2017) that minimizes cognitive load could be used to optimize the performance of judgmental model selection.

## Acknowledgements

This work is partially supported by the Beijing Natural Science Foundation (9184028). We thank Professor Nigel Harvey and Dr. George Stothart for their very insightful comments and suggestions on earlier versions of this manuscript. We also thank Ms. Ning Ding and Ms. Hua Bai for their help in conducting the experiment.

## References

Adya, M., Collopy, F., Armstrong, J. S., Kennedy, M., 2001. Automatic identification of time series features for rule-based forecasting. International Journal of Forecasting 17 (2), 143–157.

Anderson, J. R., Bothell, D., Byrne, M., Douglass, S., Lebiere, C., Qin, Y., 2004. An integrated theory of the mind. Psychological Review 111 (4), 1036–1060.

Carbone, R., Gorr, W. L., 1985. Accuracy of judgmental forecasting of time series. Decision Sciences 16 (2), 153–160.

Choi, H. H., Van Merriënboer, J. J. G., Paas, F., 2014. Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. Educational Psychology Review 26 (2), 225–244.

Collopy, F., Armstrong, J. S., 1992. Rule-Based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. Management Science 38 (10), 1394–1414.

Dirican, A. C., Göktürk, M., 2011. Psychophysiological measures of human cognitive states applied in human computer interaction. Procedia Computer Science 3, 1361–1367.

Donchin, E., 1981. Surprise! surprise? Psychophysiology 18 (5), 493–513.

Düzel, E., Cabeza, R., Picton, T. W., Yonelinas, A. P., Scheich, H., Heinze, H. J., Tulving, E., 1999. Task-related and item-related brain processes of memory retrieval. Proceedings of the National Academy of Sciences of the United States of America 96 (4), 1794.

Eggleton, I. R. C., 1982. Intuitive Time-Series extrapolation. Journal of Accounting Research 20 (1), 68–102.

Fildes, R., Goodwin, P., Lawrence, M., Nikolopoulos, K., 2009. Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. International Journal of Forecasting 25 (1), 3–23.

Fildes, R., Petropoulos, F., 2015. Simple versus complex selection rules for forecasting many time series. Journal of Business Research 68 (8), 1692–1701.

Franses, P. H., Legerstee, R., 2011. Combining SKU-level sales forecasts from models and experts. Expert Systems with Applications 38 (3), 2365–2370.

Goodwin, P., Wright, G., 1993. Improving judgmental time series forecasting: A review of the guidance provided by research. International journal of forecasting 9 (2), 147–161.

Harvey, N., Bolger, F., Mar. 1996. Graphs versus tables: Effects of data presentation format on judgemental forecasting. International Journal of Forecasting 12 (1), 119–137.

Harvey, N., Ewart, T., West, R., Apr. 1997. Effects of data noise on statistical judgement. Thinking & Reasoning 3 (2), 111–132.

Hyndman, R. J., Khandakar, Y., 2008. Automatic time series forecasting: The forecast package for R. Journal of Statistical Software 27 (3), 1–22.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., Grose, S., 2002. A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting 18 (3), 439–454.

Kenning, P., Plassmann, H., 2005. Neuroeconomics: An overview from an economic perspective. Brain Research Bulletin 67 (5), 343–354.

Lawrence, M., Goodwin, P., O'Connor, M., Önkal, D., Jan. 2006. Judgmental forecasting: A review of progress over the last 25 years. International Journal of Forecasting 22 (3), 493–518.

Lawrence, M. J., Edmundson, R. H., O'Connor, M. J., 1985. An examination of the accuracy of judgmental extrapolation of time series. International Journal of Forecasting 1 (1), 25–35.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., Simmons, L. F., 1993. The m2-competition: A real-time judgmentally based forecasting study. International Journal of Forecasting 9 (1), 5–22.

Makridakis, S., Hibon, M., 2000. The M3-Competition: Results, conclusions and implications. International Journal of Forecasting 16 (4), 451–476.

O'Connor, M., Remus, W., Griggs, K., Aug. 1993. Judgemental forecasting in times of change. International Journal of Forecasting 9 (2), 163–172.

O'Connor, M., Remus, W., Griggs, K., 1997. Going upgoing down: How good are people at forecasting trends and changes in trends? Journal of Forecasting 16 (3), 165–176.

Petropoulos, F., Fildes, R., Goodwin, P., 2016. Do big losses in judgmental adjustments to statistical forecasts affect experts behaviour? European Journal of Operational Research 249 (3), 842–852.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., Siemsen, E., May 2018. Judgmental selection of forecasting models. Journal of Operations Management 60, 34–46.

Reimers, S., Harvey, N., Oct. 2011. Sensitivity to autocorrelation in judgmental time series forecasting. International Journal of Forecasting 27 (4), 1196–1214.

Sanders, N. R., 1992. Accuracy of judgmental forecasts: A comparison. Omega 20 (3), 353–364.

Sanders, N. R., Ritzman, L. P., Jan. 1992. The need for contextual and technical knowledge in judgmental forecasting. Journal of Behavioral Decision Making 5 (1), 39–52.

Thomson, M. E., Pollock, A. C., Gönül, M. S., Önkal, D., 2013. Effects of trend strength and direction on performance and consistency in judgmental exchange rate forecasting. International Journal of Forecasting 29 (2), 337–353.

Trapero, J. R., Pedregal, D. J., Fildes, R., Kourentzes, N., 2013. Analysis of judgmental adjustments in the presence of promotions. International Journal of Forecasting 29 (2), 234–243.

Tulving, E., Kapur, S., Craik, F., Moscovitch, M., 1994. Hemispheric encoding/retrieval asymmetry in episodic memory: Positron emission tomography studies. Proceedings of National Academy of Science 91, 2016–2020.

Valeriani, D., Cinel, C., Poli, R., 2017. Group augmentation in realistic Visual-Search decisions via a hybrid Brain-Computer interface. Scientific reports 7 (1), 7772.

Wang, Y., Cui, L., Wang, H., Tian, S., Zhang, X., 2004. The sequential processing of visual feature conjunction mismatches in the human brain. Psychophysiology 41 (1), 21.

Wang, Y., Kong, J., Tang, X., Zhuang, D., Li, S., 2000. Event-related potential N270 is elicited by mental conflict processing in human brain. Neuroscience Letters 293 (1), 17–20.

## Appendix A. Forecasting models

We denote:

$\alpha$: smoothing parameter for the level ($0 \leq \alpha \leq 1$).

$\beta$: smoothing parameter for the trend ($0 \leq \beta \leq 1$).

$y_t$: actual (observed) value at period $t$.

$l_t$: smoothed level at the end of period $t$.

$b_t$: smoothed trend at the end of period $t$.

$h$: forecast horizon.

$f_{t+h}$: forecast for $h$ periods ahead from origin $t$.

SES is expressed as:

$$l_t = \alpha y_t + (1-\alpha) l_{t-1}, \tag{A.1}$$

$$f_{t+h} = l_t. \tag{A.2}$$

HES is expressed as:

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}), \tag{A.3}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta) b_{t-1}, \tag{A.4}$$

$$f_{t+h} = l_t + h b_t. \tag{A.5}$$